Robustesse et portabilités multilingue et multi-domaines des systèmes de compréhension de la parole : les corpus du projet PORTMEDIA

Fabrice Lefèvre[1], Djamel Mostefa[2], Laurent Besacier[3], Yannick Estève[4], Matthieu Quignard[5], Nathalie Camelin[4], Benoit Favre[6], Bassam Jabaian[1,3], Lina Rojas-Barahona[5]

(1) Université d'Avignon, LIA-CERI, France, {fabrice.lefevre,bassam.jabaian}@univ-avignon.fr
(2) Evaluation and Language resources Distribution Agency, France, mostefa@elda.org
(3) LIG, Grenoble, France, laurent.besacier@imag.fr
(4) LIUM, Le Mans, France, {yannick.esteve,nathalie.camelin}@univ-lemans.fr
(5) LORIA, Nancy, France, {matthieu.quignard,lina.rojas}@loria.fr
(6) LIF, Marseille, France, benoit.favre@lif.univ-mrs.fr

RÉSUMÉ

Le projet ANR PORTMEDIA avait pour objectif de compléter le corpus MEDIA afin de favoriser le développement de méthodes performantes, notamment statistiques, pour la compréhension automatique de la parole dans le cadre des systèmes de dialogues homme-machines. Les principaux axes traités sont : la robustesse aux erreurs de reconnaissance de la parole, la portabilité multilingue, la portabilité multi-domaines et la représentation sémantique haut-niveau. Ainsi tout en élaborant au sein du projet des éléments de solution à ces problématiques nous nous sommes principalement attachés à élaborer des données et meta-données permettant ensuite à d'autres groupes de recherche d'évaluer dans les meilleures conditions possibles leurs propres propositions.

ABSTRACT

Robustness and portability of spoken language understanding systems among languages and domains : the PORTMEDIA project

The ANR PORTMEDIA projet aimed at complementing the MEDIA corpus so as to foster the development of new performing approaches, including statistical approaches, for the automatic spoken language understanding in the framework of human-machine spoken dialogue systems. The main topics for which work has been carried out are : robustness to speech recognition errors, language portability, domain portability and high-level semantic representation. Thus while elaborating some solutions to theses issues inside the projet itself, we focused our efforts towards collecting new data and metadata which could help other research groups to evaluate their own propositions in the best conditions possible.

MOTS-CLÉS : corpus de dialogue oraux, compréhension de la parole, reconnaissance de la parole, multilinguisme, portabilité, représentation sémantique.

KEYWORDS: spoken language understanding, dialogue systems, speech recognition system, multilingual, portability, semantic representation.

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1: JEP, pages 779–786,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Avec le développement rapide des communications homme-machine (centres d'appels, services téléphoniques, smartphones...), la compréhension automatique de la parole (CAP) a reçu un intérêt croissant ces dernières années. Les systèmes de CAP ont été déployés dans des applications industrielles mais avec des effets mitigés jusqu'à présent. Tout d'abord les systèmes de CAP sont généralement disponibles dans une seule langue et ne supportent donc pas le multilinguisme. Ensuite les services existants utilisant des composants de CAP sont très contraints et limités par la tâche ou le domaine d'application. Enfin la qualité de l'interaction utilisateur/système est toujours loin d'être aisée et naturelle. Le projet PORTMEDIA tente d'apporter des solutions à ces difficultés en développant de nouveaux corpus visant trois objectifs distincts mais complémentaires :

– **Robustesse des systèmes de CAP aux erreurs de reconnaissance.** Les erreurs dues à la reconnaissance automatique de la parole doivent être prises en compte dans le processus de compréhension et pour cela des transcriptions automatiques doivent être mises à disposition avec les données d'apprentissage.
– **Portabilités multilingue et multi-domaines.** Les systèmes de CAP sont très dépendants de la langue et du domaine. Adapter un système à un nouveau domaine ou une nouvelle langue requiert habituellement la collecte très coûteuse d'une nouvelle grande base de données de dialogues du langage ou domaine visé et un effort important de développement. PORTMEDIA vise à permettre l'évaluation de la généricité et de la portabilité de nouvelles approches pour les systèmes de CAP.
– **Représentation sémantique haut-niveau.** Une représentation sémantique haut-niveau (High-level Semantics, HLS) est nécessaire pour prendre en compte le processus de composition sémantique intervenant au sein et entre des interactions successives de l'utilisateur.

Le projet PORTMEDIA (2009-12) est une suite du projet MEDIA (2003-07) durant lequel un corpus de 1258 dialogues en français pour *le domaine touristique* a été produit. Les partenaires du projet sont : l'université d'Avignon, ELDA, le LIG, le LIUM et le LORIA.

2 Les corpus de dialogues oraux du projet PORTMEDIA

Les systèmes de CAP de l'état-de-l'art reposent sur des modèles statistiques qui doivent être entraînés à l'aide de grand corpus de dialogues. Or, il existe très peu de corpus de dialogues homme-machine disponibles publiquement. En fait, contrairement aux autres types de parole, comme la parole lue ou les émissions télédiffusées, les seuls corpus disponibles chez LDC [1] par exemple sont des dialogues humain-humain (CallHome, CallFriends, Fisher...). Ce manque de données peut s'expliquer par la difficulté à collecter de tels corpus.

En effet, il est nécessaire de mettre au point un premier système de CAP pour collecter les données à l'aide d'un système de dialogue opérationnel. Afin de pallier cette difficulté, il est aussi possible de simuler la machine par un protocole de Magicien d'Oz (Wizard-of-Oz, WOZ) dans lequel un agent humain remplace la machine tout en tentant d'en reproduire le comportement (afin d'assurer le réalisme des données). Une fois les données collectées, des meta-données doivent être ajoutées. Alors que la transcription orthographique est une tâche bien définie, mettre au point un protocole d'annotation sémantique est bien plus complexe et requiert beaucoup

1. Linguistic Data Consortium

d'expertise. Dans ces conditions, il paraît naturel que les plus grands corpus de dialogues aient été développés par l'industrie des télécoms à l'aide de prototypes ou de systèmes déployés, comme chez AT&T *how may I help you ?* (Gorin *et al.*, 1997) ou chez France-Télécom.

PORTMEDIA a produit 2 nouveaux corpus spécifiques pour étudier la portabilité des systèmes de CAP à travers domaines et langues. Le premier corpus est composé de 604 dialogues en italien toujours sur le domaine touristique. Le second comprend 700 dialogues en français pour la réservation de billets de spectacles dans le cadre du Festival d'Avignon 2010. Les statistiques complètes sont reprises dans le tableau 1.

2.1 PM-LANG : le corpus PORTMEDIA en italien

La base de données en italien, nommée PM-LANG, a été enregistrée, transcrite et annotée en suivant les mêmes spécifications et configurations que le corpus MÉDIA initial. La seule différence entre les corpus MEDIA et PM-LANG est donc la langue parlée par les locuteurs. En 2004, 250 scénarii avaient été utilisés pour la collecte de MEDIA et une plateforme d'enregistrement téléphonique mise au point. La plateforme inclut un générateur automatique (textuel) de phrases afin d'aider les agents (compères) dans leurs réponses. Pour la base de données italienne, les mêmes outils, protocoles, scénarii et contraintes ont été retenus pour la collecte des dialogues. La seule adaptation a été de traduire les messages du WoZ et les scénarii du français vers l'italien, mais aucun changement n'a été opéré sur le contenu des scénarii (y compris les entités nommées qui ont été conservées telles quelles, par exemple les noms de lieux, d'hôtels...). Le protocole d'enregistrement est complètement décrit dans (Devillers *et al.*, 2004). La procédure d'annotation et les recommandations sont décrites dans (Maynard *et al.*, 2005). La base de données résultante est un corpus de 604 dialogues transcrits et annotés sémantiquement.

2.2 PM-DOM : le corpus PORTMEDIA en français sur un nouveau domaine

Afin d'étudier de nouvelles techniques pour la portabilité entre domaines, nous avons développé un corpus de dialogues homme-machine en français (PM-DOM) en suivant le même paradigme et les spécifications de MEDIA mais sur un domaine différent. Alors que MEDIA s'intéressait au domaine de l'information touristique, PM-DOM vise le domaine de la réservation de billets pour le Festival d'Avignon 2010. Nous avons tenté de rester aussi proche que possible des spécifications, outils et paradigmes de MEDIA et de minimiser les différences entre les 2 corpus (autre que le domaine, ce qui implique déjà une grande variabilité intrinsèque bien sûr). La seule adaptation a été de créer de nouveaux scénarii pour les appelants, d'adapter le système de gestion du dialogue pour le compère et de développer une ontologie du domaine pour l'annotation sémantique. Ce corpus comprend 700 dialogues avec transcriptions orthographiques et annotations sémantiques.

3 Pré-transcriptions et pré-annotations sémantiques des corpus PM-LANG et PM-DOM

Les transcriptions et annotations sémantiques de PM-LANG et PM-DOM ont été réalisées de manière semi-automatique. Grâce à la disponibilité du corpus MEDIA, le LIUM a développé un

reconnaisseur de parole performant pour transcrire le corpus PM-DOM. Une fois les données transcrites, elles ont été corrigées par un humain. Les corrections ont été retournées au LIUM afin de ré-entraîner les modèles et d'améliorer le système de reconnaissance de parole. Puis un nouveau lot de données était transcrit automatiquement, manuellement corrigé et retourné pour l'amélioration du système. Les détails sur le système de reconnaissance de la parole du LIUM peuvent être trouvés dans la section 3.1.

Pour les annotations sémantiques, le LIA et le LIG ont pré-annoté les deux corpus automatiquement. La pré-annotation a été validée manuellement par deux linguistes pour chacune des langues (français pour PM-DOM et italien pour PM-LANG. Le même processus itératif que pour les transcriptions a été activé. Les détails sur les modules de CAP utilisés par le LIA et le LIG sont donnés dans la section 3.2.

Nom	Lang	Domaine	#Dial	#heures	#mots	#seg concepts
MEDIA	fr	information touristique	1258	71	438k	53k
PM-DOM	fr	réservation de billet	700	40.5	293k	18k
PM-LANG	it	information touristique	604	50	218k	20k

TABLE 1 – Statistiques pour les corpus MEDIA, PM-LANG and PM-DOM.

3.1 Reconnaisseur de la parole spontanée en français

Le système de reconnaissance du LIUM pour PORTMEDIA est un système à 5 passes basé sur le système open-source SPHINX (versions 3 et 4), similaire au système LIUM'08 français décrit dans (Deléglise *et al.*, 2009) : la première passe utilise des modèles acoustiques génériques et un modèle de langage 3-grammes. Les meilleures hypothèses générées par la première passe sont utilisées pour estimer une transformation CMLLR pour chaque locuteur. Utilisant des modèles acoustiques SAT et MPE et les transformations CMLLR, la deuxième passe génère des graphes de mots. Dans la troisième passe, les graphes de mots sont re-scorés en utilisant un score acoustique inter-mots plus performant. La passe suivante re-calcule les scores des graphes de mots avec un 4-grammes. Enfin, la dernière passe génère un réseau de confusion dont est extraite l'hypothèse finale par la méthode du décodage par consensus.

Les modèles acoustiques ont été estimés sur les corpus ESTER-1 (Galliano *et al.*, 2005), ESTER 2 (Galliano *et al.*, 2009) et EPAC (Estève *et al.*, 2010) : l'ensemble représentant environ 280 heures d'émissions radio-télédiffusées. Les modèles de langage et le vocabulaire ont été extraits directement du corpus MEDIA (conformément aux résultats de (Lefèvre *et al.*, 2005) sur l'apprentissage multi-source). Afin de traiter le premier lot de données enregistrées de PM-DOM, le système utilise un vocabulaire de 5k mots et des 4-grammes ont été appris sur l'ensemble d'apprentissage de MEDIA. Le taux d'erreur mots de ce système sur le test MEDIA était de 25,2% sur les énoncés des appelants uniquement (*i.e.* sans prendre en compte les tours de parole des agents qui sont généralement mieux reconnus car très formatés). Le tableau 2 montre les taux d'erreurs atteints par les versions successives du système de reconnaissance après chaque itération du processus de pré-transcription présenté dans la section 3. Quatre lots de données ont été traités automatiquement. À chaque itération, le vocabulaire et les modèles de langage étaient mis à jour d'après les corrections manuelles.

Itération	Dialogues	Global	WoZ	Appelants
0 (init)	**1-100**	**46,9%**	41,3%	53,2%
1	**101-300**	**15,9%**	7,4%	39,5%
2	**301-500**	**15,8%**	6,9%	37,2%
3	**501-700**	**15,9%**	8,2%	35,6%

TABLE 2 – Taux d'erreur mots de la pré-transcription automatique pour chaque itération. Les phrases du WoZ et de l'appelant sont calculées séparément.

3.2 Les systèmes de CAP pour le français et l'italien

Afin de pouvoir réaliser la pré-annotation sémantique pour les corpus français et italien, nous avons utilisé un étiqueteur CAP basé sur une méthode statistique : les champs conditionnels markoviens (Conditional Random Fields, CRFs). Un corpus sémantiquement annoté est nécessaire pour entraîner un tel système. Pour le corpus français PM-DOM, des modèles furent entraînés directement sur les données MEDIA et de nouvelles entités nommées ont été ajoutées simultanément afin de prendre en compte les nouveautés dans les données. La pré-annotation est une combinaison des sorties proposées par les systèmes de CAP et de détection d'entités nommées.

Comme il n'existe pas de corpus équivalent pour la langue cible du corpus PM-LANG, nous avons proposé de porter automatiquement le corpus MEDIA français en italien. Plusieurs approches pour la portabilité d'un système de CAP entre langues ont été étudiées et évaluées (Jabaian *et al.*, 2010; Lefevre *et al.*, 2010) et la meilleure a été appliquée pour créer un nouveau corpus annoté. L'approche retenue consiste à traduire automatiquement le corpus MEDIA français en italien puis de porter l'annotation sémantique sur les données italiennes. La traduction a été réalisée par un système de traduction automatique statistique à base de segments sous-phrastiques (Phrase-based Statistical Machine Translation, PB-SMT) entraîné sur un corpus parallèle (obtenu en traduisant manuellement un sous-ensemble des données françaises). Le transfert de l'annotation est basé sur un alignement automatique entre les phrases françaises et italiennes. En d'autres termes, la méthode consiste en la projection des concepts à l'aide de l'alignement des corpus source et cible. Dans la mesure où le corpus français était déjà annoté de façon segmentale, nous avons proposé d'utiliser directement l'information de l'alignement mot-à-mot. Pour ce faire nous avons développé un algorithme qui utilise les informations d'alignement et de segmentation : à chaque segment conceptuel en français, l'algorithme associe les mots correspondants en italien en se référant à l'alignement. Cette stratégie a permis d'annoter l'ensemble du corpus traduit en italien (y compris la partie traduite manuellement). Le corpus italien permet alors d'entraîner un étiqueteur sémantique qui à son tour permet de réaliser une première itération de pré-annotation en italien. Pour les itérations suivantes, la correction manuelle d'un premier lot est ajoutée au corpus d'entraînement et les modèles réappris.

Une évaluation des performances du modèle de CAP italien est décrite dans le tableau 3. Seuls sont reportés les taux d'erreur en concept avec ou sans utilisation des corpus PM-LANG et MEDIA traduit pour l'entrainement des modèles (les taux mesurés sur le test MEDIA traduit manuellement sont donnés à titre de référence). Le meilleur résultat sur le test PM-LANG 17,6% est obtenu par une combinaison des deux corpus. Pour le corpus PM-DOM, le modèle entrainé avec les données du corpus obtient un CER de 19,1%.

Apprentissage	Test	Sub	Del	Ins	CER
MEDIA	MEDIA	3,1	15,0	2,3	20,5
	PM-LANG	3,8	13,9	3,1	20,8
PM-LANG	MEDIA	4,7	17,4	3,2	25,3
	PM-LANG	3,6	12,1	3,3	18,9
MEDIA **et** PM-LANG	MEDIA	2,8	14,6	2,1	19,5
	PM-LANG	3,9	9,0	4,6	17,6

TABLE 3 – Évaluation (CER %) des modeles de CAP italiens en fonction de l'ensemble d'apprentissage.

3.3 Gains de productivité

Durant la transcription et l'annotation sémantique, nous avons régulièrement comparé les gains dus aux pré-transcriptions et annotations semi-automatiques. Dans cette optique, le protocole suivant a été implémenté. Lors de chaque itération un ensemble de 10 dialogues était transcrit (resp. annoté) par deux annotateurs différents. Le premier réalisait la transcription (resp. annotation) à partir de la pré-transcription (resp. pré-annotation) tandis que le second annotateur réalisait les mêmes opérations sans hypothèses initiales. Les gains en productivité mesurés sont reportés dans la figure 1.

Pour les transcriptions, on observe que le temps mis à transcrire un dialogue est divisé par deux avec l'utilisation des transcriptions automatiques. Pour l'annotation sémantique, les gains de productivité sont de plus de 50% pour l'italien et 40% pour le français.

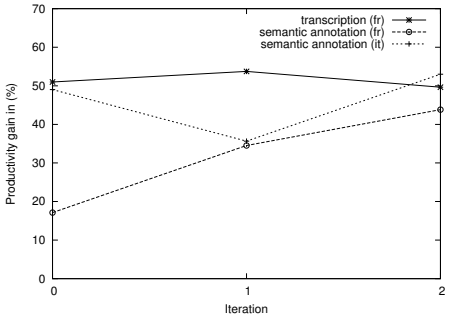

FIGURE 1 – Gains de productivité (en %) pour la transcription et l'annotation sémantique des 2 corpus PM-LANG et PM-DOM.

4 Annotation sémantique haut-niveau

L'utilisation d'une sémantique haut-niveau (High Level Semantic, HLS), une représentation sémantique hiérarchique, a été étudiée pour l'annotation du corpus MEDIA. À cet effet, nous nous sommes référés au langage pour les interfaces multimodales (MultiModal Interface Language, MMIL) pour générer des structures guidées par l'ontologie du domaine qui supportent les informations linguistiques utiles de la syntaxe jusqu'au discours.

Ainsi, les traits fins des actes de dialogues, prédicats et arguments sont correctement définis pour les énoncés par rapport à l'annotation conceptuelle séquentielle déjà disponible pour le corpus. Toutefois, cette annotation représente un véritable challenge. Pour commencer, nous avons traité les énoncés les plus complexes, contenant des prédicats et arguments se chevauchant et des énoncés elliptiques contenant des prédicats et/ou des arguments implicites. Nous avons élaboré le guide d'annotation et annoté manuellement un sous-ensemble d'énoncés supposés être représentatifs des aspects les plus complexes de l'annotation HLS, en terme de constituants (Rojas-Barahona *et al.*, 2011), à l'aide d'un outil graphique développé spécifiquement. Une architecture incrémentale a ensuite été élaborée pour l'annotation semi-automatique du corpus complet, ainsi que les moyens d'évaluer l'annotation fournie (Rojas-Barahona et Quignard, 2011).

Cette annotation est en cours de correction par des annotateurs experts afin de fournir un *gold standard* pour l'ensemble du corpus. Celui-ci servira à implémenter et tester des méthodes d'apprentissage supervisé pour l'annotation automatique de cette sémantique complexe.

5 Conclusion

Dans cet article, nous avons présenté les différents apports du projet PORTMEDIA visant à favoriser le développement de méthodes statistiques pour la compréhension de la parole. Principalement, le travail a pris la forme de la mise à disposition de corpus permettant une évaluation pertinente et aisée des méthodes étudiées. Ainsi, 4 axes principaux ont été couverts :

– robustesse aux erreurs de reconnaissance : mise à disposition de transcriptions automatiques avec un système à l'état de l'art des données
– portabilité multilingue : collecte d'un nouveau corpus en italien, comprenant un ensemble de test et un ensemble d'adaptation ;
– portabilité multi-domaine : collecte d'un nouveau corpus sur un nouveau domaine (réservation de billets), avec un ensemble de test et d'adaptation ;
– annotation sémantique hiérarchique : élaboration d'une représentation sémantique de haut-niveau permettant la prise en compte d'informations au niveau de la phrase complète (les spécifications sont prêtes, l'annotation du corpus MEDIA complet est en cours de finition).

Enfin une série d'évaluations a été réalisée sur les nouvelles données fournies permettant notamment de vérifier leur qualité. Ces évaluations seront poursuivies avec la recherche de collaboration externe au projet afin de fournir les données avec une référence de performance. L'ensemble du corpus ainsi réalisé rejoindra le catalogue d'ELDA [2] dans le courant de l'année.

2. http ://catalog.elra.info/

Remerciements

Ce travail a été partiellement financé par l'Agence Nationale pour la Recherche : projet PORTMEDIA ANR 08 CORD 026 01. Plus d'informations sur www.port-media.org.

Références

DELÉGLISE, P., ESTÈVE, Y., MEIGNIER, S. et MERLIN, T. (2009). Improvements to the LIUM french ASR system based on CMU Sphinx : what helps to significantly reduce the word error rate ? *In Interspeech 2009*, Brighton (United Kingdom).

DEVILLERS, L., MAYNARD, H., ROSSET, S., PAROUBEK, P., MCTAIT, K., MOSTEFA, D., CHOUKRI, K., CHARNAY, L., BOUSQUET, C., VIGOUROUX, N., (5), F. B., ROMARY, L., ANTOINE, J., VILLANEAU, J., VERGNES, M. et GOULIAN, J. (2004). The french media/evalda project : the evaluation of the understanding capability of spoken language dialogue systems.

ESTÈVE, Y., BAZILLON, T., ANTOINE, J., BÉCHET, F. et FARINAS, J. (2010). The EPAC corpus : manual and automatic annotations of conversational speech in french broadcast news. *In Proceedings of the Seventh conference on International Language Resources and Evaluation, Valletta, Malta*.

GALLIANO, S., GEOFFROIS, E., MOSTEFA, D., CHOUKRI, K., BONASTRE, J. F. et GRAVIER, G. (2005). The ESTER phase II evaluation campaign for the rich transcription of French broadcast news. *In EUROSPEECH-05*, volume 1, pages 1149–1152, Lisbonne, Portugal.

GALLIANO, S., GRAVIER, G. et CHAUBARD, L. (2009). The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. *In In Interspeech 2009*.

GORIN, A., RICCARDI, G. et WRIGHT, J. (1997). How may i help you. *Speech Communication*, 23:113–127.

JABAIAN, B., BESACIER, L. et LEFEVRE, F. (2010). Investigating multiple approaches for slu portability to a new language. *In Proceedings of the 11th Annual Conference of the International Speech Communication Association, Japan, Septembre 2010*.

LEFÈVRE, F., GAUVAIN, J.-L. et LAMEL, L. (2005). Genericity and portability for task-independent speech recognition. *Computer Speech & Language*, 19(3):345–363.

LEFEVRE, F., MAIRESSE, F. et YOUNG, S. (2010). Cross-lingual spoken language understanding from unaligned data using discriminative classification models and machine translation. *In Proceedings of the 11th Annual Conference of the International Speech Communication Association, Japan, Septembre 2010*.

MAYNARD, H., ROSSET, S., AYACHE, C., KUHN, A. et MOSTEFA, D. (2005). Semantic annotation of the media corpus for spoken dialog. pages 3457–3460.

ROJAS-BARAHONA, L. M., BAZILLON, T., QUIGNARD, M. et LEFEVRE, F. (2011). Using mmil for the high level semantic annotation of the french media dialogue corpus. *In In : Proceedings of the 9th International Conference on Computational Semantics, Oxford, January 2011*.

ROJAS-BARAHONA, L. M. et QUIGNARD, M. (2011). An incremental architecture for the semantic annotation of dialogue corpora with high-level structures. a case of study for the media corpus. *In Proceedings of the SIGDIAL 2011 Conference*, page 332–334, Portland, Oregon. Association for Computational Linguistics, Association for Computational Linguistics.